How much can Behavioral Targeting Help Online Advertising?

Jun Yan[1], Ning Liu[1], Gang Wang[1], Wen Zhang[2], Yun Jiang[3], Zheng Chen[1]
[1]Microsoft Research Asia
Beijing, 100080, China
{junyan, ningl, gawa, zhengc}@microsoft.com
[2] Department of Automation
University of Science & Technology
Hefei, 230027, China
v-wenzha@microsoft.com
[3]ShangHai Jiao Tong University
Shanghai, 200240, China
v-yunjiang@microsoft.com

ABSTRACT
Behavioral Targeting (BT) is a technique used by online advertisers to increase the effectiveness of their campaigns, and is playing an increasingly important role in the online advertising market. However, it is underexplored in academia how much BT can truly help online advertising in search engines. In this paper we provide an empirical study on the click-through log of advertisements collected from a commercial search engine. From the experiment results over a period of seven days, we draw three important conclusions: (1) Users who clicked the same ad will truly have similar behaviors on the Web; (2) Click-Through Rate (CTR) of an ad can be averagely improved as high as 670% by properly segmenting users for behavioral targeted advertising in a sponsored search; (3) Using short term user behaviors to represent users is more effective than using long term user behaviors for BT. We conducted statistical t-test which verified that all conclusions drawn in the paper are statistically significant. To the best of our knowledge, this work is the first empirical study for BT on the click-through log of real world ads.

Categories and Subject Descriptors
I.6.4 [**Computing Methodologies**]: Simulation and Modeling – *model validation and analysis*. E.0 [**Data**]: General

General Terms
Measurement, Performance, Economics, Experimentation.

Keywords
User segmentation, online advertising, Behavioral Targeting (BT), Click-Through Rate (CTR).

1. INTRODUCTION
With the rapid growth of the World Wide Web (WWW), online advertising channels, such as sponsored search [4], contextual ads [1], and Behavioral Targeting (BT), are showing great market potentials. However, in contrast to the widely studied general sponsored search, BT, which refers to the delivery of ads to targeted users based on information collected on each individual user's web search and browsing behaviors, is still underexplored in academia. To encourage more research on BT and possibly to further develop this market, we provide an empirical study on the click-through log of advertisements collected from a commercial search engine to seek the answer to the question: how much can BT help online advertising?

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

We use 7 days' ads click-through log data coming from a commercial search engine, dated from June 1st to 7th 2008, to compare different BT strategies and validate the effectiveness of BT. The log dataset records all users' search click behavior, which includes both Web page clicks and ad clicks of users. To be clear of any privacy concerns, we did not study any user demographic and geographic information for the targeted advertising. In order to answer the question of how much BT can help online advertising, we split our investigation into answering three questions step by step:

First of all, we aim to empirically answer the question of ***whether BT truly has the ability to help online advertising***. Note the assumption behind BT is that the users who have similar search or browsing behaviors will have similar interests and thus have higher probability to click the same ad than the users who have different online behaviors. If this assumption is true, online users can be grouped into different user segments according to their behaviors for targeted ads delivery. Thus our first mission in this work is to validate whether the BT assumption is true. It is the foundation of further BT exploration. In this paper, we propose two novel measures, i.e., the within- and between- ads user similarities. These similarity measures help us understand whether the users who clicked the same ad will be more similar than the users who clicked different ads.

Secondly, we aim to answer the question of ***how much BT can help online advertising using commonly used evaluation metrics***. The performance of online advertising is commonly measured by ads Click-Through Rate (CTR) or the revenue from advertisers. In this work, we propose to observe how much BT can improve ads CTR through the segmentation of users into a number of small groups for targeted ads delivery. We did not use the revenue as the evaluation metric since the information for ads revenue is not readily available for research purpose. We show that ads CTR can be significantly improved through utilizing BT technologies compared with traditional ads delivery without behavioral targeting. In order to confirm CTR improvements are significant, the statistical paired t-test is applied to the results of all ads we collected. The t-test values, which are expected to be less than 0.05, validate the statistical significance of our experiment results. The answer to this question can confirm our conclusion drawn from the first question.

Finally, we aim to answer the question of ***which BT strategy can work better than others for ads delivery***. According to the definition of "Behavioral Targeting (BT)" [15], there are two strategies to represent the users' behavior, i.e., Web browsing behavior and search behavior, which can be denoted by users' clicked pages or search queries respectively. In this paper, we formally compare these two BT strategies for user segmentation. The results provide valuable guidelines on how to implement

behavioral targeted advertising in real world applications. In addition, to make the conclusions more convincing, we use ads click entropy, precision, recall and F-measure for comparing different BT strategies.

As a summary, from all experiments conducted in this paper, we can draw our conclusions in three steps.

1. Through verifying the basic assumption of BT by defining the within- and between- ads user similarities, we observe that the users who clicked the same ad can be over 90 times more similar than the users who clicked different ads. From this observation we can conclude that BT can truly help online advertising by segmenting users based on user behaviors for ads delivery.
2. Through studying ads CTR before and after user segmentation for ads delivery, we observe that ads CTR can be improved by as much as 670% over all the ads we collected. The t-test results, which are very close to zero, confirm the statistical significance of CTR improvements. In addition, we notice that if we can further design more advanced BT strategies, such as novel user representation approaches and novel user segmentation algorithms, ads CTR can be further improved beyond 1,000%.
3. Through comparing different user representation strategies for BT, we draw the conclusion that the user search behavior, i.e. user search queries, can perform several times better than user browsing behavior, i.e., user clicked pages. Moreover, only tracking the short term user behaviors are more effective than tracking the long term user behaviors, for targeted ads delivery.

The rest of this paper is organized as follows. In Section 2, we introduce some background about BT and discuss the different BT strategies to be validated and compared in this paper. In Section 3, we introduce the dataset to be used in this study. In Section 4, we summarize our experiment configuration including our proposed evaluation metrics. In Section 5, we show our observations from the experiment results. Finally in Section 6 we introduce our conclusion and future work.

2. BEHAVIORAL TARGETING

Among various online advertising techniques, Behavioral Targeting has been attracted much attention recently. According to the definition of "Behavioral Targeting" in Wikipedia [15], which is a good summary of BT related articles, "BT uses information collected on an individual's web-browsing behavior, such as the pages they have visited or the searches they have made, to select which advertisements to display to that individual. Practitioners believe this helps them deliver their online advertisements to the users who are most likely to be influenced by them." In our problem configuration, each individual is defined as a search user. According to this definition, BT is generally used for improving the influence of online advertising by targeting the most relevant user for the ads being displayed and vice versa. There are generally two steps in behavioral targeted advertising: user segmentation and user segments ranking. The first step aims to segment users according to their behaviors and the second step aims to rank targeted user segments for an advertisement. Thus all the user segmentation strategies to be studied in this paper will not depend on any specific query.

Recently, there have been a large number of commercial systems proposed for targeted advertising. For instance, Yahoo! smart ads [22] collects around 169M registered users for behavioral targeting, which also integrates the demographic and geographic targeting. Adlink [11] uses the short user session during search for behavioral targeted advertising. DoubleClick [17] additionally utilizes some extra features such as browse type and the operating system of users for user segmentation. Specificmedia [12] proposes to assign a score for predicting the interest and purchase intent of each targeted user. Besides these, there are many other popularly used commercial BT systems such as TACODA [21], Revenue Science [20], Phorm [19], Blue Lithium [14], Almond Net [13], NebuAd [18], Burst [16], etc. Though an increasing number of commercial BT systems appeared, they have no public works in academia to answer the question of how much BT can truly help online advertising in commercial search engines. In this paper, we explore this problem in three steps, which can be summarized as three sub-questions,

1. Does BT truly have the ability to help online advertising? To answer this question, we validate the basic assumption of BT, i.e. whether the users who clicked the same ad always have similar browsing and search behaviors and the users who clicked different ads have relatively different Web behaviors.
2. How much can BT help online advertising using commonly used evaluation metrics? To answer this question, we use the difference between ads CTR before and after applying BT strategies as the measurement, i.e. the degree of CTR improvement is considered as a measurement of how much BT can help online advertising. The statistical t-test is utilized to secure the significance of our experiment results.
3. What BT strategy works better than others for ads delivery? We consider two types of BT strategies, which are (1) represent user behaviors by users' clicked pages and (2) represent user behaviors by users' search queries respectively. In addition, how long the user behaviors have occurred in the log data is also considered for user representation. Detailed configurations of different BT strategies are introduced in the remaining part of this section.

To represent user behavior by their page-views, we consider the clicked URLs of search users as their profiles. In other words, all the users can be considered as a user-by-URL matrix, where each row of this matrix is a user and each column of this matrix is a URL. We adopt the classical Term Frequency Inverse Document Frequency (TFIDF) indexing [8] by considering each user as a document and considering each URL as a term for mathematical user representation. Mathematically, all users are represented by a real valued matrix $U \in R^{g \times l}$, where g is the total number of users and l is the total number of URLs that have been clicked in our dataset. A user is a row of U, which is a real valued vector with the weight for each entry to be,

$$u_{ij} = (log(\#times\ user\ i\ clicked\ URL\ j) + 1) \times log \frac{l}{\#\text{user clicked URL } j}$$

where $i = 1,2, \dots g, j = 1,2, \dots l$.

On the other hand, we also build the user behavioral profile by simply considering all terms that appear in a user's queries as his previous behaviors. Thus we can represent each user in the Bag of Words (BOW) model [8] where each term is considered as a feature. We use Porter Stemming [3] to stem terms and then remove stop words and terms which only appeared once in a user's query texts. Consequently, 470,712 terms are removed and the remaining 294,208 terms reserved. After this preprocessing, each user is represented by BOW with corresponding term frequency. We use the same TFIDF [8] indexing as the one used for building the user-by-URL matrix to index the users by query terms. To avoid the bias of the data, the query, using which a user

clicked an ad, will be discounted when we calculate the term frequency in representing this user. For example, if a user u used query q three times and she clicked ad a once, then we consider this user to have the behavior q only twice. All users can then be represented by a real valued matrix. Mathematically, all users are represented by a user-by-query matrix, without loss of generality, we use the same symbol $U \in R^{g \times h}$ to represent this matrix, where g is the total number of users and h is the total number of terms that have appeared in user queries. A user is a row of U, which is a real valued vector. Both ways allow us to represent users as numerical vectors, thus the similarity between users can be easily calculated in the vector space.

Different commercial BT systems consider different time windows for tracking the user interests. Many commercial systems consider relatively long term user behaviors while others prefer to consider the short term user behaviors for BT. However, no previous evidence is shown to prove which strategies are better. In this work, we consider the long term user behavior and the short term user behavior as two different BT strategies respectively. As a preliminary study, we use 1 day's user behavior as their short term profile and use 7 days' user behavior as their long term behavioral profile in our experiment configuration. As a summary, we will validate and compare four different BT strategies in this paper. They are,

1. **LP**: using **L**ong term user behavior all through the seven days and representing the user behavior by **P**age-views;
2. **LQ**: using **L**ong term user behavior all through the seven days and representing the user behavior by **Q**uery terms;
3. **SP**: using **S**hort term user behavior (1 day) and representing user behavior by **P**age-views;
4. **SQ**: using **S**hort term user behavior (1 day) and representing user behavior by **Q**uery terms.

All the experiments in this paper will be conducted according to these four different user representation strategies respectively.

3. DATASET

In this section, we mainly introduce the dataset to be used in this study. It integrates a sponsored search click-through log with general purpose search click-through log, which comes from the same commercial search engine. In other words, the log dataset records all users' search click behavior, which contains both web page clicking and ad clicking. We use 7 days' click-through log data ranging from June 1st to 7th 2008. To identify the unique users, we utilize the user IDs in the log data. The IDs are assigned according to the cookies of users stored in their operating systems. To be clear of any privacy concerns, no other user information, such as demographic and geographic, are logged or predicted. The detailed data format is summarized in Table 1, where a synthetic example is given for demonstration instead of a real ads click record. The last column of Table 1 is the explanation description.

In order to draw convincing conclusions, we filter out robots from our log data before conducting the experiment. For example, some user IDs may have up to thousands of clicks within one day, which are explicit online robots. To filter them out through simple heuristic rules, we set an upper threshold of user clicks to be 100 per day. As a result, anyone who has more than 100 clicks a day will be removed. In addition, we only deal with English queries in this paper. Finally, the remaining qualified data contains 6,426,633 unique users and 335,170 unique ads within the seven days. We filter out all the ads that have less than 30 clicks within these seven days, since they cannot be used to draw reliable statistical conclusions. Overall, we have 17,901 ads remaining for this study. The experiment results in this paper are averaged over these 17,901 advertisements.

Table 1. Format of click-through log used in our study.

UserID	UID030608473X	A user ID for each unique user.
QueryText	xbox	The detailed query text used by the user
QueryTime	08-06-03 21:15:47	The time when the query was issued
ClickTime	08-06-03 21:16:02	The time when the click occurred after the query was issued
ClickURL	http://www.xbox365.com	The URL which has been clicked by the user
IsAd	0	A Boolean value to show the clicked URL is an ad or not
NumberAd	3	The number of ads displayed in the search results
DisplayAd	http://video-games.half.ebay.com/ http://accessories.us.dell.com/ http://www.gamefly.com	The URL list of all the ads that displayed by the query. (To save space, we only reserve top domain of the ad URL in this example.)

4. EXPERIMENT CONFIGURATION

To answer the three questions listed in Section 2, we systematically explore the BT problem by a set of experiments on real world ads click-through log. In Section 4.1, we introduce the mathematical symbols, which will be used throughout the experiments, with detailed experiment configurations. In Section 4.2, we propose the evaluation metrics we will use in this study.

4.1 Symbols and Experiment Setup

Before showing the detailed experiment configuration, we first define some mathematical symbols, which will be used throughout the experiments. Let $A = \{a_1, a_2, \ldots a_n\}$ be the set of the n advertisements in our dataset. For each ad a_i, suppose $Q_i = \{q_{i1}, q_{i2}, \ldots q_{in_i}\}$ are all the queries which have displayed or clicked a_i. Through these queries, we can collect all the corresponding users who have displayed or clicked a_i. Suppose the group of users who have either displayed or clicked a_i is represented by $U_i = \{u_{i1}, u_{i2}, \ldots u_{im_i}\}$. We define a Boolean function,

$$\delta(u_{ij}) = \begin{cases} 1 & \text{if } u_{ij} \text{clicked } a_i \\ 0 & \text{otherwise} \end{cases}$$

to show whether the user u_{ij} has clicked ad a_i.

BT aims to group users into segments of similar behaviors and deliver different ads to different groups of users. In this work, we used two common clustering algorithms, k-means [10] and CLUTO [7] for user segmentation. Suppose the users are segmented into K segments according to their behaviors. We use the function,

$$G(U_i) = \{g_1(U_i), g_2(U_i), \ldots, g_K(U_i)\},\ i=1,2,\ldots n$$

to represent the distribution of U_i under a given user segmentation results, where $g_k(U_i)$ stands for all the users in U_i

who are grouped into the k^{th} user segment. Thus the k^{th} user segment can be represented by,

$$g_{k=} \bigcup_{i=1,2,\ldots,n} g_k(U_i)$$

As a summary of key steps in the experiment, we first represent the users by their behaviors using different types of BT strategies, which are introduced in Section 2. After that, we group the users according to their behaviors by the commonly used clustering algorithms. Finally, we evaluate how much BT can help online advertising by delivering ads to good user segments. To provide convincing evaluation results for the performance of different BT strategies, we provide the evaluation metrics from different perspectives in the next subsection.

4.2 Evaluation Metrics

In this subsection, we introduce the evaluation metrics for different BT strategies. They are, within- and between- ads user similarity, improvement of ads Click-Through Rate (CTR), ads click Entropy and F-measure. We additionally utilize the paired t-test to verify the statistical significance of our experiments. The evaluation metrics are organized step by step to answer how much BT can truly help online advertising.

4.2.1 Within- and Between- Ads User Similarity

A basic assumption of BT is that the users who have similar search or browsing behavior will have similar interests and thus have a higher probability to click the same ad than the users who have different online behaviors. Our first measurement aims to validate this assumption to see whether BT has the potential to help online advertising. Suppose the similarity between a pair of users u_{ij} and u_{st} is $Sim(u_{ij}, u_{st})$. If the assumption of BT is true, the similarity between users who clicked the same ad must be larger than the similarity between users who clicked different ads. As introduced in Section 2, we have already represented all users in the numerical vector space. Thus the classical Cosine similarity can be utilized for the similarity computation between users. Without loss of generality, we use the same symbol u_{ij} to represent both users and the vector representation of his user behavior. The similarity between users is defined as,

$$Sim(u_{ij}, u_{st}) = \frac{< u_{ij}, u_{st} >}{||u_{ij}||||u_{st}||}$$

where $<\ ,\ >$ stands for the vector inner-product and $||.||$ is the vector 2-norm. For ad a_i, the user similarity, who clicked it, is defined as the within ads user similarity,

$$S_w(a_i) = \frac{2}{l_i(l_i-1)} \sum_{\delta(u_{ij})=1} \sum_{\substack{\delta(u_{it})=1 \\ t \neq j}} Sim(u_{ij}, u_{it})$$

where $l_i = \sum_j \delta(u_{ij})$ is the number of users who clicked ad a_i. $S_w(a_i)$ shows how similar the users are, who clicked the same ad according to their behaviors. We are also interested in how similar the users are who clicked different ads. We define the between ads user similarity as,

$$S_b(a_i, a_s) = \frac{1}{l_i l_s} \sum_{\delta(u_{ij})=1} \sum_{\delta(u_{st})=1} Sim(u_{ij}, u_{st})$$

It describes how similar the users are, who clicked ad a_i and ad a_s respectively. We further define a ratio between $S_w(a_i)$ and $S_b(a_i, a_s)$ as,

$$R(a_i, a_s) = \frac{S_w(a_i) + S_w(a_s)}{2S_b(a_i, a_s)}$$

Intuitively a large R score means the two ads have a large within ads similarity and small between ads similarity. The larger the $R(a_i, a_s)$ is, the more confident we are on the basic assumption of BT for a pair of ads a_i and a_s.

4.2.2 Ads Click-Through Rate

If we have validated the basic assumption of BT, a further question is how much BT can help online advertising. The performance of online advertising is generally measured by the ads CTR or revenue. Since it is hard for us to track the revenue of all advertisers for research purposes, we propose to observe whether BT can improve ads CTR. The CTR of ad a_i is defined as the number of users who clicked it over the number of users who either clicked it or only displayed it, i.e.

$$CTR(a_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \delta(u_{ij})$$

After user segmentation, the CTR of a_i over user segment g_k is,

$$CTR(a_i|g_k) = \frac{1}{|g_k(U_i)|} \sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})$$

where $|g_k(U_i)|$ is the number of users in $g_k(U_i)$. If there exist some user segments where the CTR of the same ad can be significantly improved in contrast to the CTR without user segmentation, then we say BT is valuable for online advertising.

4.2.3 F-measure

Even though we can validate the effectiveness of BT by ads CTR, it is not sufficient to draw convincing conclusions. For example, if we observe that there has a user segment g_k, which satisfies that $CTR(a_i|g_k) > CTR(a_i)$, it can only provide evidence that there has a segment of users who are more interested in ad a_i than other users. It cannot guarantee we have segmented as many users as possible, who potentially will click a_i. In other words, the improvement of CTR after user segmentation can only validate the precision of BT strategies in finding potentially interested users. The recall is not guaranteed. Motivated by this, we propose to adopt the classical F-measure [6] for BT evaluation. If we consider the users who clicked a_i as positive instances and consider the users who are displayed ad a_i but did not click it as negative instances, the Precision and Recall are defined as,

$$Pre(a_i|g_k) = CTR(a_i|g_k)$$

$$Rec(a_i|g_k) = \frac{\sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})}{\sum_{j=1}^{m_i} \delta(u_{ij})}$$

It can be seen that the larger the precision is, the more accurate we can segment the clickers of a_i. The larger the recall is, the better the coverage we can achieve in collecting all the clickers of a_i through user segmentation. To integrate these two parts, we propose to utilize the classical F-measure for results evaluation,

$$F(a_i|g_k) = \frac{2Pre(a_i|g_k)Rec(a_i|g_k)}{Pre(a_i|g_k) + Rec(a_i|g_k)}$$

The larger the F measure is, the better the performance we can state to have achieved by user segmentation for BT. Note the F-measure is not only used to evaluate a single user segment, it can be used to evaluate a group of selected user segments if we allow delivering one ad to multiple user segments.

4.2.4 Ads Click Entropy

Intuitively, if the clickers of an ad a_i dominate some user segments and seldom appear in other user segments, we can easily deliver our targeted ads to them by selecting the segments they dominated. However, suppose the clickers of a_i are uniformly distributed in all user segments, if we aim to deliver the targeted ads to more interested users, we have to deliver the ad to more users who are not interested in this ad simultaneously. Motivated by this, we further define the ads click Entropy to show the effectiveness of different BT strategies. For ad a_i, the probability of users in segment g_k, who will click this ad, is estimated by,

$$P(g_k|a_i) = \frac{1}{m_i} \sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})$$

According to the mathematical formulation of Entropy, given G, we define the ads click Entropy of ad a_i as,

$$Enp(a_i) = -\sum_{k=1}^{K} P(g_k|a_i) log P(g_k|a_i)$$

Thus the larger the Entropy is, the more uniformly the users, who clicked ad a_i, distribute among all the user segments. The smaller the Entropy is, the better results we will achieve.

4.2.5 Summary

All the evaluation metrics introduced in this section are used to evaluate each independent ad separately. One way for global evaluation over all the ads is to observe the average performance. However, the average results cannot guarantee the improvements to be statistically significant. Some occasionally big improvement may lead to the improvement of average results. Thus in this work, we propose to consider the paired t-test [5] to guarantee the statistical significance of the results. For t-test, we compare two types of experiment configurations. The statistical t-test is conducted on the comparison of results over all the ads.

5. BT RESULTS

In this section, we present our experiment results for validating and comparing different BT strategies. In Section 5.1, we validate the basic assumption of BT to show its potential in helping online advertising. In Section 5.2, we experimentally show how much BT can improve ads CTR. In Section 5.3, we give some more evaluated results by the ads click Entropy and F-measure. After that in Section 5.4, we discuss some strategies to further improve BT performance. Finally in Section 5.5, we summarize our observations.

5.1 Assumption of BT

We use the within- and between- ads similarity of users to validate whether the users who clicked the same ad may have similar behaviors and the users who clicked different ads will have relatively different behaviors. Let

$$S_w = \sum_i S_w(a_i)/n \text{ and } S_b = \sum_i \sum_s S_b(a_i, a_s)/n^2$$

be the average within ads and average between ads user similarity over all ads of our collected dataset respectively. In addition, the averaged ratio can be calculated by,

$$R = \sum_i \sum_s R(a_i, a_s)/n^2.$$

Table 2 gives detailed results, which are averaged over our ads collection. Note each row of Table 2 stands for a user representation strategy for BT. In order to make the experiment fair, all the queries that led to the ad clicks are removed from the user representation when calculating LQ and SQ.

Table 2. Within- and between- ads user similarity.

	S_w	S_b	R
LP	0.1417	0.0252	28.9217
LQ	0.2239	0.0196	44.2908
SP	0.1532	0.0281	24.5086
SQ	0.2594	0.0161	91.1890

From the results of Table 2, we can observe that the average S_w is larger than the average S_b no matter which BT strategy we use. This means that the users who clicked the same ad are more similar than those who clicked different ads according to their behaviors. The most significant one is SQ with the average R as large as 91.189 compared with other BT strategies. This means the within ads similarity of users, which are represented by their short term search behaviors, can be around 90 times larger than the corresponding between ads similarity. Among all the ads we collected in our dataset, about 99.37% pairs of ads have the property that $R(a_i, a_s) > 1$, which means that for most of the ads, the within ads user similarity is larger than the between ads user similarity. This table also tells us that the search queries will be more effective than clicked pages for user representation in BT. In addition, only tracking the short term user behaviors for BT may give a better performance than tracking long term user behaviors.

To validate whether the difference between S_w and S_b is statistically significant, we implement the paired t-test to compare the results of S_w with that of S_b. Table 3 shows the t-test results of different BT strategies, which are all less than 0.05. This table accurately validates the observation that, statistically, the within ads user similarity is always larger than the between ads similarity.

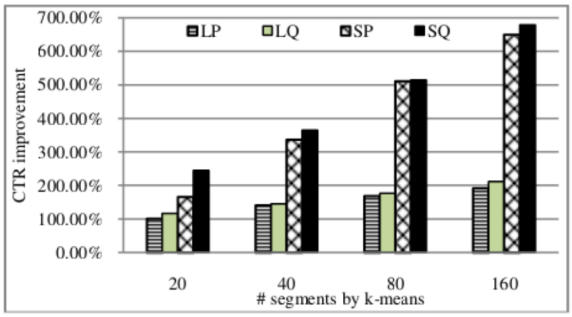
Table 3. T-test for S_w against S_b.

	LP	LQ	SP	SQ
T-test	4.1E-294	0	3.3E-282	0

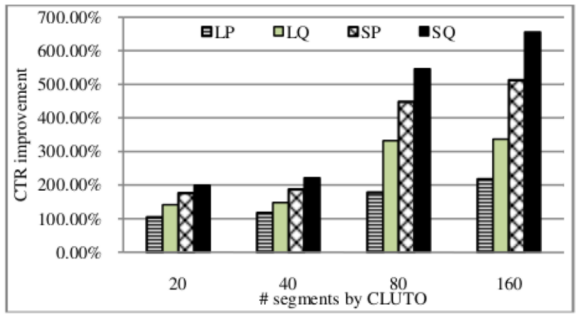
As a summary, the experiment results in this section tell us that the users who clicked the same ad will have more similar behaviors than the users who clicked different ads. This verified the basic assumption of BT and motivates us to segment users according to their behaviors for targeted advertising.

5.2 BT for Online Advertising

In this section, we aim to answer on how much BT can help online advertising in terms of ads CTR. As introduced in Section 4.1, we firstly represent users by their behavior under different BT strategies. Then we group the similar users into segments according to their behavior. Here both k-means and CLUTO are used for user clustering. Finally, we look at the clustering results to see whether there are any user segments that can significantly improve the CTR of given ads. We group all our users into 20, 40, 80 and 160 clusters no matter which clustering algorithm is used. For each ad, we can calculate $CTR(a_i)$ over all users. We can also calculate its CTR over different user segments, i.e. $CTR(a_i|g_k)$. Let $g_*(a_i) = argmax\{CTR(a_i|g_k), k = 1,2, \dots, K\}$, then $g_*(a_i)$ is a user segment that have the highest CTR for a_i. Note $g_*(a_i)$ is only optimal in terms of ads CTR, it is not guaranteed to have the largest number of impressions for ad a_i. In our future work, we will study how to select the user segments which have both high CTR and high impressions. In this study, we use,

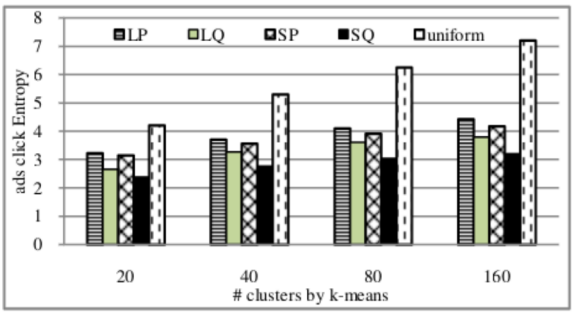


(a) User clustering by k-means

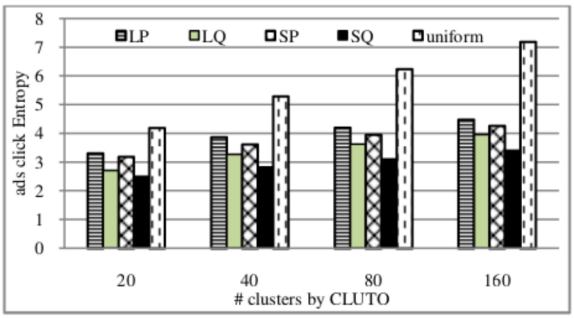


(b) User clustering by CLUTO

Figure 1. CTR improvements by user segmentation for BT.



(a) Cluster by k-means



(b) Cluster by CLUTO

Figure 2. Ads click Entropy of user segmentation for BT.

$$\Delta(a_i) = \frac{CTR(a_i|g_*(a_i)) - CTR(a_i)}{CTR(a_i)}$$

represent the CTR improvement degree of ad a_i by user segmentation in BT. As preliminary results, this ratio is used to reflect how much BT can help online advertising. We show the average results in Figure 1, i.e. improvement$\Delta = \sum_i \Delta(a_i)/n$.

From the results show in Figure 1, we can observe that in most of the cases, SQ gives the best CTR improvements in contrast to the CTR without targeted ads delivery. The ads CTR can be improved as high as 670% by the simple user segmentation strategies used for behavioral targeted advertising. Note different clustering algorithms, which are k-means and CLUTO, will not change our conclusions much. The ads CTR can be improved further by segmenting the users into more user segments. Two interesting observations are (1) using short term user behavior such as SQ and SP can achieve better CTR than using long term user behavior for user representation; and (2) for user representation, the search queries issued by users can always perform a little bit better than the pages visited by users. These two observations are both statistically significant through paired t-test (<0.05).

We provide some analysis for the conclusions drawn from Figure 1 through manual case studies. To our understanding, the reason why the short term user behaviors are more effective than the long term user behaviors for targeted advertising is that the users have multiple interests that always change rapidly. If we use relatively long term user behaviors to group users, many users who have lost old interests or have more interests to other topics are grouped together. On the other hand, the short term user interests can well reflect the users' recent and focused interests, which can significantly improve ads CTR. As for the reason why the search queries can work a little bit better than the page clicking for BT, we found that in the dataset we analyzed, the queries have a strong correlation to the ads displaying while the page clicks have no strong correlation to that. Thus a reason which leads to this difference may be the bias of the data. Another reason is that the users' interests can be directly reflected by search queries. However, the pages clicked by users are not always a reflection of user interests, since with search engines, not all users know what they will see before they click a page.

To further show that the improvements of ads CTR are significant over all ads, we compare the CTR before and after user segmentation through paired t-test. The results are displayed in Table 4. From the small t-test results in Table 4, which are all less than 0.05, we can draw the conclusion that the improvements of CTR by BT are statistically significant. The query based user representation is more significant than the page click based user representation.

Table 4. T-test of CTR improvements by BT

T-test	LP	LQ	SP	SQ
K-means	3.79E-4	5.02E-18	4.36E-5	0
CLUTO	6.95E-4	3.66E-16	8.87E-5	0

As a summary, the experiment results in this section tell us that through user segmentation, the behavioral targeted advertising can significantly improve ads CTR if we deliver our ad to some proper segments of users. The average CTR improvement rate can be as high as 670% by using proper user representation and user segmentation algorithms, where using the short term user search behavior for user representation, i.e. SQ, can perform better than the others, which are LP, LQ and SP.

Table 5. F measure of different BT strategies

		LP	LQ	SP	SQ
K-means (20 segments)	***Pre***	8.67%	8.60%	13.35%	**17.08**%
	Rec	10.20%	22.34%	7.63%	**25.58**%
	F	0.08	0.10	0.08	**0.16**
CLUTO (20 segments)	***Pre***	8.62%	8.56%	14.61%	**19.13**%
	Rec	10.01%	20.51%	7.86%	**21.43**%
	F	0.08	0.10	0.07	**0.15**
K-means (40 segments)	***Pre***	8.84%	9.23%	19.76%	**20.53**%
	Rec	9.48%	18.20%	4.83%	**20.75**%
	F	0.08	0.10	0.06	**0.16**
CLUTO (40 segments)	***Pre***	8.76%	9.14%	19.38%	**22.80**%
	Rec	8.44%	**17.88**%	4.52%	17.78%
	F	0.08	0.10	0.06	**0.14**
K-means (80 segments)	***Pre***	9.02%	9.63%	23.47%	**23.49**%
	Rec	8.93%	17.62%	4.06%	**19.35**%
	F	0.08	0.10	0.06	**0.16**
CLUTO (80 segments)	***Pre***	8.85%	9.51%	23.09%	**27.00**%
	Rec	7.82%	**16.65**%	4.00%	15.55%
	F	0.07	0.10	0.06	**0.15**
K-means (160 segments)	***Pre***	9.09%	9.93%	25.68%	**25.81**%
	Rec	8.54%	17.98%	3.92%	**19.78**%
	F	0.074	0.10	0.06	**0.17**
CLUTO (160 segments)	***Pre***	8.87%	9.84%	25.43%	**31.02**%
	Rec	7.24%	**15.58**%	3.78%	14.52%
	F	0.07	0.10	0.06	**0.15**

5.3 More Evaluation

Although the experiment results in Section 5.2 have shown that BT can significantly help online advertising in terms of CTR, it can only guarantee the precision of the top user segments in delivering targeted ads. It cannot reflect how the clickers of an ad distribute across all the user segments. As another evaluation metric, we show ads click Entropy results in Figure 2. According to the definition of Entropy, the smaller the ads click Entropy is, the better the performance we are expected to have achieved. The ideal case is that all the users who have clicked the same ad distribute in the same user segment. We can then easily deliver this ad to the targeted user segment. The ads click Entropy for this ideal case should be zero. On the other hand, suppose we have K user segments, the worst case is the users who clicked the same ad uniformly distribute in all the K user segments. If this bad condition occurs, the ads click Entropy will get its maximum value, which is,

$$Enp(a_i) = -\sum_{k=1}^{K} P(g_k|a_i) log P(g_k|a_i)$$
$$= -\sum_{k=1}^{K} \frac{1}{K} log \frac{1}{K} = logK$$

Under this condition, it is impossible for us to deliver ad a_i to a good user segment for targeted advertising. In our experiment, we segment the users into 20, 40, 80 and 160 segments respectively by using two different clustering algorithms. Thus the worst Entropy results for them are 4.32, 5.32, 6.32 and 7.32 respectively. We represent the worst cases by "uniform" in Figure 2.

From Figure 2, we see that the ads click Entropy of different strategies is always smaller than its counterpart of the uniform distribution. The best performance is always achieved by SQ no matter if the users are segmented by k-means or CLUTO. Different from the observations through ads CTR, the runner up among the four BT strategies is LQ, followed by SP and LP. The ranked orders of the four strategies in ads click Entropy are different from their counterparts in ads CTR, since they evaluate the results from two different perspectives. One is used to measure how many users who displayed an ad will click it within a user segment, which is also known as the precision of the user segment. The other is used to measure how the clickers of an ad distribute across all the user segments, which is similar to the averaged recall of user segments. However, no matter which measurement we use, we draw the same conclusion that SQ gives the best performance.

To formally study the tradeoff between precision (*Pre*) and recall (*Rec*) for user segmentation in BT, in this section we adopt the F-measure, which integrates both precision and recall for evaluation. Table 5 shows the averaged F measure (*F*) over all the advertisements in our data collection. From this table, we can see that in most of the cases, SQ gives the best performance. Though LQ sometimes can give good recall, the F measure will be consistently worse than with SQ. As a summary, in terms of the F measure, using queries to represent users can always perform better than using clicked pages; using short term user behaviors for user representation can outperform using the long term behaviors.

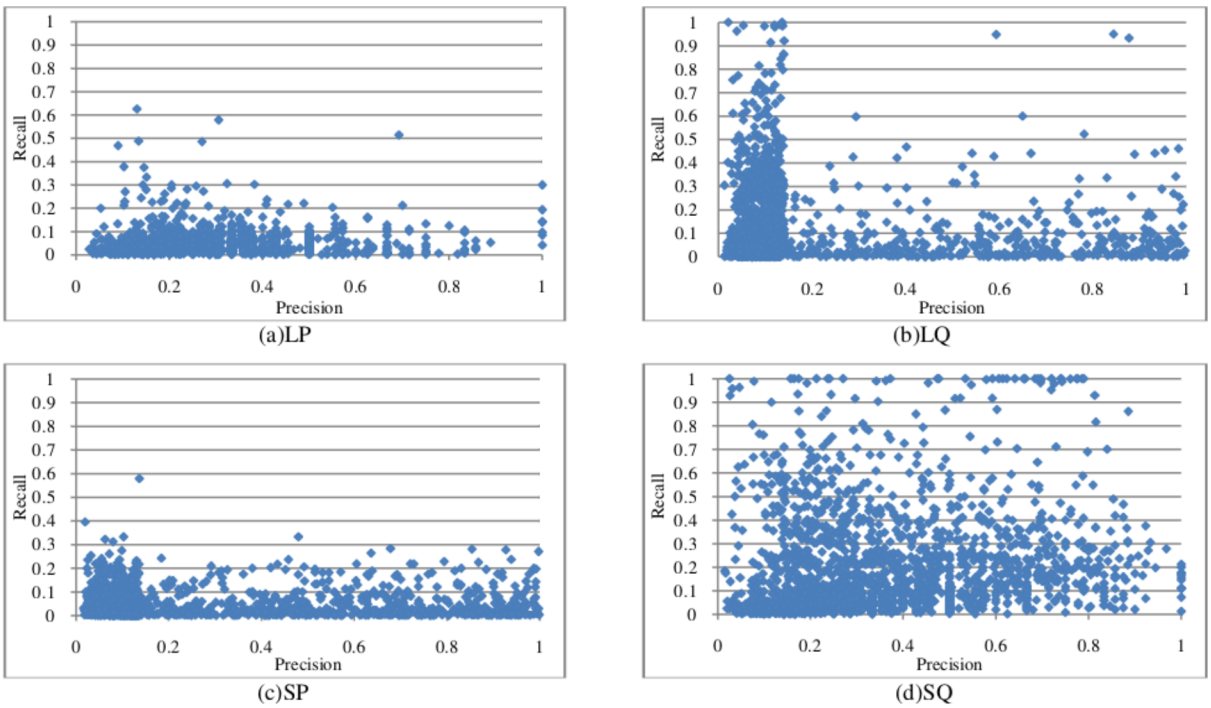
Besides Table 5, to provide more detailed results about BT performance, we present the scatter plot of the precision and recall over all the ads we collected in Figure 3. We only plot the P-R scatter for the CLUTO algorithm since all our experiments above show that the k-means and CLUTO can provide us similar observations. In addition, to save space, only the results for the 160 user segments are provided. The x-axis of the figure stands for precision while the y-axis stands for recall and thus each ad can be represented as a point in the two dimensional space. To make the figure clearer, we only randomly sampled 3,000 ads, which can reflect the real data distribution, to plot in the figure.

From Figure 3, we can clearly see the differences among the four user representation strategies. LP and SP have very limited recall. LQ may have higher recall but the precision is limited for most of the ads. SQ has relatively larger number of ads that have both high precision and high recall. This confirms that averagely SQ will give the best performance for BT. As a summary, in this section we evaluated the performance of different user representation strategies for BT by ads click Entropy and F measure respectively. From all the experiments, we can draw the same conclusion that SQ is the best among the four strategies we studied. Users represented by their search queries can work better than representing users by their clicked Web pages for user segmentation. The short term user behaviors are more effective than long term user behaviors in representing the users' interests.

5.4 Further Discussion

All the experiments presented in previous sub-sections are implemented under the fundamental experiment configurations. For example, we only used the search queries and clicked pages respectively as the user behavior profiles to denote the users in vector space. On the other hand, we only considered the k-means clustering and CLUTO for user segmentation in the numerical vector space. It is unclear for us whether some better user representation strategies or better machine learning algorithms for user segmentation can provide better BT performance. Although the exploration on advanced BT algorithms is out of the scope of this paper, to encourage BT related research, we give some preliminary experiment results to show that BT performance has much more potential to be further improved through developing advanced algorithms.

We firstly explore some other user representation strategies for BT. For instance, one way to better represent the users is to integrate their search behavior with their browsing behavior. In other words, we can combine their search queries and clicked pages for user representation. Another way is to consider the user search sessions instead of considering their search behaviors independently for user representation. In other words, we can use the continuous search queries as a stream to represent users. In this subsection, we discuss the combination of user search and browsing behaviors for user representation. Figure 4 gives some interesting observations, which is measured by ads CTR.



(a)LP (b)LQ (c)SP (d)SQ

Figure 3. Scatter plot of Precision and Recall over all the ads (CLUTO-160 user segments).

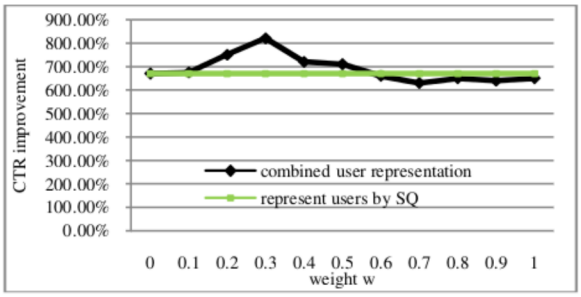


Figure 4. CTR improvements with combined user representation (k-means for user segmentation)

In this figure, we directly combine the vector representation of users, which are queries (Q) and clicked pages (P) respectively, by a weighted longer vector. For simplicity, we only consider the short term representation (S) in this set of experiments. Suppose the vector representation of a user u in SQ is $u_{SQ} \in R^h$, the vector representation of the same user in SP is $u_{SP} \in R^l$, then the combined user representation is formatted as

$$(wu_{SQ}, (1-w)u_{SP}) \in R^{h+l}$$

which is a longer weighted vector. $w, 0 \leq w \leq 1$, is the weight, which stands for the x-axis in Figure 4. We also involved the best performance we have achieved in Section 5.2 as baseline, i.e. the results of SQ, for comparison.

From this figure, we can see that through changing the user representation strategies, it is possible to further improve ads CTR. Using the weight 0.3 can give the best performance in this example. We believe some other advanced user modeling algorithms, such as the Hidden Markov Model and Maximum Entropy Model, can provide better performance for BT. However, more exploration is out of the scope of this paper. Here, we use this example to show that some quite simple algorithm designing work can give further improvement for the performance of BT. Note in Figure 4 we only use k-means for user segmentation, since we observe from the results of previous subsections that the general clustering algorithms, such as k-means and CLUTO, provide similar conclusions to us.

In our future experiments, we will answer how BT performance can be improved through investigating other user segmentation algorithms besides the general ones used in this paper. We explore this problem by applying a modified Min-wise hashing clustering algorithm (MinHash) [9], which has been used by Google News [2] as a document clustering algorithm, for user segmentation. The results are shown in Figure 5, which is measured by CTR improvement. In this example, we still use SQ as the user representation strategy for demonstration.

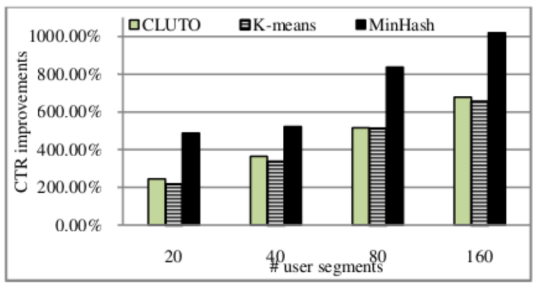


Figure 5. CTR improvements by Min-wise hashing clustering (MinHash) for user segmentation.

From this figure, we see that the CTR of MinHash clustering can be significantly improved compared with the classical k-means and CLUTO. A reason for this improvement is that this clustering algorithm is more flexible than the other two since it allows the same user to belong to multiple user segments. More exploration about the user segmentation algorithms will be introduced in our future work. Through this figure, we aim to show the fact that the performance of targeted advertising can be significantly improved through designing better user segmentation algorithms.

In this work, we only selected one user segment for each ad when we calculated the precision and recall. In real world applications, we can rank several user segments for an ad, which can improve the recall. Thus it is an interesting problem worthy of further study that after user segmentation, how to select user segments for ads delivery? Instead of proposing novel user segments ranking algorithms, in this section, we use an example to show how the precision and recall will change if we select more than one user segments for ads delivery. In this example, we use CLUTO as the user segmentation algorithm and segment the users into 160 segments. SQ is used as the instance for study. Figure 6 shows the change of precision and recall with the change of user segments number we considered for ads delivery.

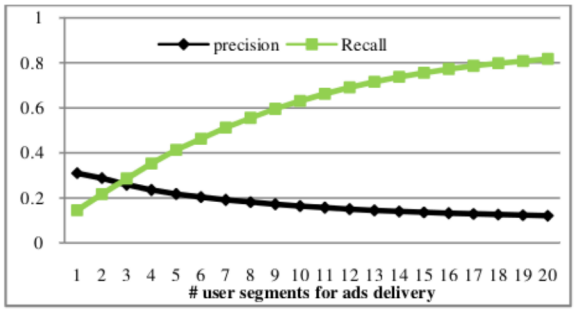


Figure 6. Change of precision and recall with increase number of user segments for ads delivery.

This figure tells us that if we deliver each ad to more user segments, the precision of ads delivery will be decreased slightly while the recall with be increased significantly. Thus we need to balance the precision and recall in real world application by selecting a proper number of user segments for ads delivery. More detailed user segment ranking algorithms will be discussed in our future work. As a summary, in this subsection, we empirically verified that it is possible for us to further improve BT performance through designing better user interests modeling and user segmentation algorithms. This presents an underexplored research direction to the online advertising research community. In addition, after user segmentation, we propose to explore user segment ranking algorithms for targeted ads delivery.

5.5 Results Summarization

In this subsection, we review all the experiment results of this paper. Firstly, in Section 5.1, we validated that the users who clicked the same ad will be more similar than the users who clicked different ads according to their behaviors. The ratio, which is the within ads user similarity over the between ads user similarity can be as high as 91. This verifies the basic assumption of user segmentation for BT. After that, in Section 5.2, we showed that if we segment the users according to their behaviors by some classical clustering algorithms, ads CTR can be improved by as much as 670% by selecting the proper user segments for ads

delivery. To compare different user representation strategies for BT, in Section 5.3, we use ads click Entropy and F measure as evaluation metrics. From the results we draw the conclusion that SQ can always give the best performance. Using queries for user representation is better than using the user clicked pages. Using the short term user behaviors for user representation is better than using the long term user behaviors. Finally in Section 5.4, we showed that through designing some advanced user representation and user segmentation algorithms, ads CTR can be further improved to more than 1,000%. This set of experiments in Section 5.4 introduces several directions in BT research such as the user segments ranking problem, which requires further research.

6. CONCLUSION AND FUTURE WORK

In this work, we provide a systematic study on the ads click-through log of a commercial search engine to validate and compare different BT strategies for online advertising. To our best knowledge, this work is the first systematic study for BT on real world ads click-through log in academia. Through experiments on the log with more than 6 million search users and 17,901 real world ads, we draw the conclusions that (1) the users who clicked the same ad will be more similar than the users who clicked different ads; (2) ads CTR can be averagely improved as high as 670% over all the ads we collected if we directly adopt the most fundamental user clustering algorithms for BT; and (3) for the user representation strategies, which are defined in the definition of BT, tracking the short term user search behavior can perform better than tracking the long term user browsing behavior. These three conclusions can answer the three questions: whether the basic assumption of BT is true, how much BT can help online advertising, and which BT strategy can perform better in the behavioral targeted advertising. We believe this study can provide valuable guidelines for the behavioral targeted advertising research and related system design.

In our future work, we will conduct more studies along several directions step by step. As introduced in Section 5.4, some advanced user segmentation algorithms can give better results in behavioral targeted advertising. We will explore the detailed BT algorithms for further improving the online ads influence. User behavioral data is always of large scale and incremental. In terms of computation, we will mainly study the algorithms that can deal with large scale user data and the rapidly changing user behavior data stream. In addition, user behavior modeling is underexplored for BT. We will study better user representation strategies such as user search sessions, the content of user clicked pages and user browsing trials for targeted advertising. Finally, after the users are segmented, how the user segments can be ranked for a given ad is an important problem but is not deeply studied in this report. We will also study the user segment ranking problem for behavioral targeting in our future work.

7. ACKNOWLEDGMENTS

The authors would like to thank all reviewers for their valuable comments. The authors would like to thank Ying Li for her valuable comments, discussion and kind help in improving the manuscript.

8. REFERENCES

[1] A. Broder, M. Fontoura, V. Josifovski and L. Riedel. A semantic approach to contextual advertising. In Proceedings of SIGIR '07 (Amsterdam, July 2007), ACM Press, 559-566.

[2] A. S. Das, M. Datar, A. Garg and S. Rajaram. Google news personalization: scalable online collaborative filtering. In Proceedings of WWW '07 (Banff, May 2007), ACM Press, 271-280.

[3] C. J. van Rijsbergen, S. E. Robertson and M. F. Porter. New models in probabilistic information retrieval. British Library Research and Development Report, No. 5587, 1980.

[4] D. C. Fain and J. O. Pedersen. Sponsored search: a brief history. In Bulletin of the American Society for Information Science and Technology, 2005.

[5] D.R. Cox and D.V. Hinkley. Theoretical statistics. Chapman and Hall, London, 1974.

[6] G. Hripcsak and A.S. Rothschild. Agreement, the F-Measure, and reliability. Information Retrieval Journal of the American Medical Informatics Association, 2 (May 2005), 296-298.

[7] G. Karypis. CLUTO: a software package for clustering high-Dimensional data sets. University of Minnesota, Dept. of Computer Science.

[8] G.Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management: an International Journal, 24 (1988), 513-523.

[9] P. Indyk and R. Motwani. Approximate nearest neighbor: towards removing the curse of dimensionality. In Proceedings of the 30th Annual ACM Symposium on Theory of Computing (Dallas, May 1998), ACM Press, 604-613.

[10] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman and A. Wu. An efficient K-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell., 24 (July 2000), 881-892.

[11] Adlink https://www.google.com/adsense/login/en_US/?gsessionid=Dc28hZShnCI

[12] Specificmeida http://www.specificmedia.co.uk/

[13] Almond Net http://www.almondnet.com/

[14] Blue Lithium http://www.bluelithium.com/

[15] http://en.wikipedia.org/wiki/Behavioral_targeting

[16] Burst http://www.burstmedia.com/

[17] Double Click http://www.doubleclick.com/products/dfa/index.aspx

[18] NebuAd http://www.nebuad.com/

[19] Phorm http://www.phorm.com/

[20] Revenue Science http://www.revenuescience.com/advertisers/advertiser_solutions.asp

[21] TACODA http://www.tacoda.com/

[22] Yahoo! Smart Ads http://advertising.yahoo.com/marketing/smartads/